International Journal of Research in Advent Technology, Vol.3, No.11, November 2015
E-ISSN: 2321-9637
Available online at www.ijrat.org

An Enhanced Sentiment Variation Analysis Using Genetic and Bayesian Information Extractor Using Twitter Datasets

R.Urega[1], Dr.M.Devapriya[2]
[1]M.Phil Scholar, [2]Assistant professor,
PG & Research Department of Computer Science, Government Arts College,Coimbatore-18.
uregarajendran@gmail.com[1], devapriya_gac@rediffmail.com[2]

Abstract: As the increase of social networking, people started to share information through different kinds of social media. Twitter platform is valuable to follow the public sentiments. Knowing users point of views and reasons behind them at various point is an important study to take certain decisions. Categorization of positive and negative opinions is a process of sentiment analysis. It is very useful for people to find sentiment about the person, product etc. before they actually make opinion about them. In this paper, TTSM (Temporal Topic Sentiment Mining) along with BIE have been proposed. The system can effectively find reviews on different topics, further these find temporal segmented results for opinion detection using semantic variations. With the use of genetic algorithm, the accuracy has been increased.

Index Terms -Twitter, Public sentiment,Sentiment analysis,Temporal Topic Sentiment Mining, Bayesian Information Extractor.

1. INTRODUCTION

Extraction and analysis of information is a very tedious process. Several blog sitesand social sites are today's most popular interactive medium to communicate, share, and disseminate a considerable amount of human life information and feedback.Analyzing certain data from those networks may not be so easy and accurate [3].Users express their opinions about products or services they consume in blog posts, shopping sites, or review sites. Reviews on a wide variety of commodities are available on the Web. Sentiment, opinion and topic mining from a set of documents is the major contribution of the proposed system. Sentiment analysis on twitter data helps to expose opinions of peoples.

One important analysis is to find possible reasons behind sentiment variation, which can provide important decision making information. It is generally difficult to find the exact reason of sentiment variations. The emerging topics which are discussed in the different changing periods are connected to the some genuine reasons behind the variations. It will be consider these emerging topics as possible reasons. It defines two models. First one is TTSM (Temporal Topic Sentiment Mining) model can filter unwanted topics and then extra topics and opinion to reveal the variations of sentiment. Another model is BIE(Bayesian Information Extractor). The BIE model can find the best matching keyword in each domain in natural language to provide sentiment variations. The two proposed models were evaluated on Twitter dataset.

2. REVIEW OF LITERATURE

Machine learning techniques have been widely deployed for sentiment and opinion classification at various levels, e.g., from the document level, to the sentence and word/phrase level. On the document level, one tries to classify documents as positive, negative or neutral, based on the overall sentiments expressed by opinion holders. There are several lines of representative work at the early stage [8], [9].

Chenghua et al [8] used weakly supervised learning with mutual information to predict the overall document sentiment and opinion by averaging out the sentiment and opinion orientation of phrases within a document.

Pang et al. [9] classified the polarity of movie reviews with the traditional supervised machine learning approaches and achieved the best results using SVMs. In their subsequent work [1], the sentiment and opinion classification accuracy was further improved by employing a subjectivity detector and performing classification only on the subjective portions of reviews.

The annotated movie review data set (also known as polarity data set) used in [9] and [1] has later become a benchmark for many studies [12], [6]. Whitelaw et al. [12] used SVMs to train on combinations of different types of appraisal group features and bag-of-words features, whereas Kennedy and Inkpen [6] leveraged two main sources.

Goldman and Roy[2] explored various strategies for customizing sentiment and opinion classifiers to new domains, where training is based on a small number of labelled examples and large amounts of unlabelled in-domain data. It was found that directly applying a classifier trained on a particular domain barely outperforms the baseline for another domain.

In the same vein, more recent work [10], [4] focused on domain adaptation for sentiment and opinion classifiers. W.Jiang [10] addressed the domain transfer problem for sentiment and opinion classification using the structural correspondence learning (SCL) algorithm, where the frequent words in both source and target domains were first selected as candidate pivot features and pivots were then chosen based on the mutual information between these candidate features and the source labels. They achieved an overall improvement of 46 percent over a baseline model without adaptation. Li and Huang [4] combined multiple single classifiers trained on individual domains using SVMs.

3. RELATED WORKS

OPINE [7] is an unsupervised information extraction system which extracts fine-grained features, and associated opinions, from reviews. OPINE's use of the Web as a corpus helps identify product features with improved precision compared with previous work. OPINE uses a novel relaxation-labeling technique to determine the semantic orientation of potential opinion words in the context of the extracted product features and specific review sentences; this technique allows the system to identify customer opinions and their polarity with high precision and recall. It Need less training data and Applies semantic orientation which provides high precision and recall**.**

In this paper [11], formally define the problem of topics sentiment analysis and propose a new probabilistic topic sentiment mixture model (TSM) to solve this problem. In this model, have effectively learn general sentiment models, (1)extract topic models,(2)orthogonal to sentiments, which can represent the neutral content of a subtopic; and (3) extract topic life cycles and the associated sentiment dynamics.

In this paper [5] they presented a joint model of text and aspect ratings for extracting text to be displayed in sentiment summaries. The model uses aspect ratings to discover the corresponding topics and can thus extract fragments of text discussing these aspects without the need of annotated data.

In this paper [13], this proposed two generative models to discover aspects and sentiment in reviews. SLDA constrains that all words in a single sentence be drawn from one aspect. ASUM unifies aspects and sentiment and discovers pairs of aspect, sentiment aspects. The aspects and senti-aspects discovered from reviews of electronic devices and restaurants show that SLDA and ASUM capture important evaluative details of the reviews. ASUM is also capable of capturing aspects that are closely coupled with a sentiment.

4. PROBLEM DEFINITION

There are several challenges in analyzing the sentiment of the web user reviews. First, a word that is considered to be positive in one situation may be considered negative in another situation. Take the word "long" for instance. If a customer said a laptop's battery life was long, that would be a positive opinion. If the customer said that the laptop's start-up time was long, however, that would be a negative opinion. These differences mean that an opinion system trained to gather opinions on one type of product or product feature may not perform very well on another.

However, the more informal the medium (twitter or blogs for example), the more likely people are to combine different opinions in the same sentence. For example: "the movie bombed even though the lead actor rocked it" is easy for a human to understand, but more difficult for a computer to parse. Sometimes even other people have difficulty understanding what someone thought based on a short piece of text because it lacks context. For example, "That movie was as good as his last one" is entirely dependent on what the person expressing the opinion thought of the previous film. The recent work on sentiment variations does not currently extract non-noun opinion features due to the limitation of only considering nouns (noun phrases) for candidate feature extraction in the dependence parsing phase.

Detection and tracking of topics have been studied extensively in the area of topic detection and Mining. The main task is to either classify a

new tweet into one of the known opinion label (such as positive or negative) or to detect that it belongs to none of the known categories (Neutral). Then the structure of topics has been modelled and analyzed through dynamic model selection and temporal text mining. All the existing system had the use of textual content of the documents, but not the social content of the documents. The social content has been utilized in the study of social network (SN) sites. However, SN networks are often analyzed in a NLP setting. The current proposal lies in focusing on the social content of the documents (posts) and in combining this with a change analysis.

5. RESEARCH METHODOLOGY

Our Methodology starts with the process of learning the model from a corpus of training data and classifying the unseen data based on the trained model. In general, classification tasks are often divided into several sub-tasks.

5.1 Data Pre-Processing:

When a new message received in the Twitter, the textual content will be extracted and stored as the dataset for every data from the dataset then the content separation has been done. For example, if user A receives a message from user B Twitter, the dataset will be like below.

"Top CEOs express great satisfaction after meeting Modi"

"All the Best to Narendramodi for making a great initiative"

5.1.1 Replacing All Sequences Of Whitespace

This process has been done using simple regular expression concepts. A simple pattern matching functionality can effectively identify these types of characters. The whitespace includes characters (tabs, spaces and newline characters) by a single space.

5.1.2 Eliminating "Stop Words"

After concatenating the words, stop word elimination process begins. Stop words are a division of natural language. The motive that stop-words should be removed from a text is that they make the text look heavier and less important for analysts. Removing stop words reduces the dimensionality of term space. The most common words in text contents are prepositions, articles and pro-nouns, etc. that does not give the meaning of the documents. These words are treated as stop words. Example for stop words: *the, in, a, an, with, etc.* Stop words are removed from documents because those words are not measured as keywords in text mining applications.

5.1.3 Stemming

This method is used to identify the root/stem of a word. For example, the words connect, connected, connecting, connections all can be stemmed to the word "connect" The purpose of this method is to remove various suffixes, to reduce the number of words, to have accurately matching stems, to save time and memory space. For that, the proposed framework used porter stemmer algorithm.

5.1.4 Porter Stemmer Algorithm

The pre-processing process includes the stemming process, which eliminates unnecessary keys. All stemming algorithms can be roughly classified as affix removing, statistical and mixed. Affix removal stemmers apply set of transformation rules to each word, trying to cut off known prefixes or suffixes.

Porter stemmer utilizes suffix stripping techniques rather than prefix methods.

Step 1: Gets rid of plurals and -ed or -ing suffixes

Step 2: Turns terminal y to i when there is another vowel in the stem

Step 3: Maps double suffixes to single ones:- ization, -ational, etc.

Step 4: Deals with suffixes, -full, -ness etc.

Step 5: Takes off -ant, -ence, etc.

Step 6: Removes a final –e

The above steps represent the process and elimination of porter stemmer algorithm. The importance of the stemmer algorithm is, it reduces the difficulties of data classification when the training data's are insufficient. Thus it effectively eliminates the suffix words such as 'ed', 'ing' etc,

The pseudo code for the above algorithm is represented below,

A sample output of the stemming process is represented in the table 5.1 below.

1. String s
2. Split string s and stored into s[].
3. For each word in s[]
4. If S[i].text end with "ed"

5. Remove the two keys from the word.
6. Store s1[i].
7. Else If S[i].text end with “ing”
8. Remove the three keys from the word.
9. Store s1[i].
10. else if ends("s") || ends("ss")
11. do step 9

Table 5.1 Output of porter stemmer

Words	Processed words
Strings	String
Clusters	Cluster
Operated	Operate
Realized	Realize

6. RESULTS

TTSM compares the performance for effective sentiment variation finding on twitter.This experiment tests on twitter messages and email based dataset contains 50-60 messages.Recall and Precision to evaluate the performance of this approach. Recall is the percentage of the number of messages that have been correctly filtered over the total number of data records on a twitter. Precision is the percentage of the number of messages that have been correctly filtered over the total number of data records that have been extracted.

Pr = Cc / Ce

Rr = Cc / Cr

Where, Cc is the count of correctly extracted messages and total messages,

Ce is the count of filtered messages, and

Cr is the actual count of messages in the training samples.

The number of messages in different twitter page varies from a few to hundreds. Consequently, pages with many messages will dominate the record level metrics. To use a page-level metric, namely page- level precision defined as,

Pp =Cp /Na

Where, cp is the count of correctly filtered messages, which means that all the messages in the pages are correctly filtered and summarized to the user, Na is the count of all the pages from which messages are filtered . To assume that each input age contains at least two messages and data extraction is performed on all input pages.

As per theoretical comparison and proof from the current experiment setup, the comparison study has developed. The proposed TTSM shows better results, as a well-known data record extraction system.

Table 6.1Performance Analysis

Tweets	LDA	TTSM_BIE
Total tweets taken	350	340
Correctly classified tweets	330	335
Record level Precision %)	94.2	98.5
Record level Recall (%)	96.3	99.7

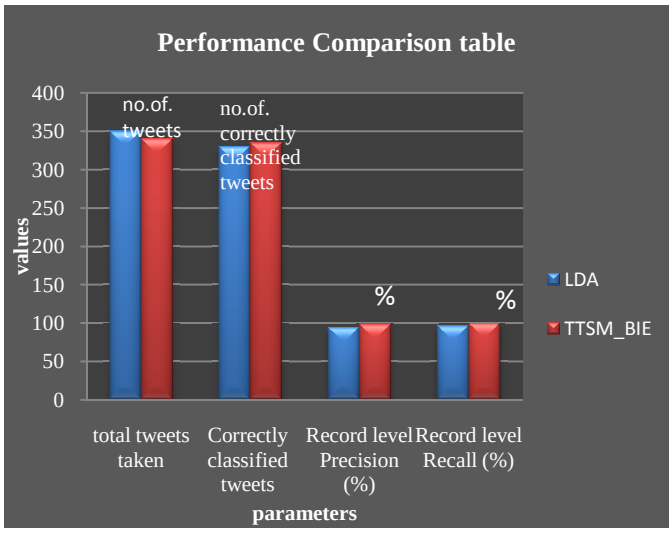

Figure 6.1 Performance Comparison Table

The above chart shows the graphical representation comparing the performance of TTSM with the existing LDA approach. As per Table 6.1, this approach has much better experimental results than existing approach LDA, and in almost every domain this approach significantly outperforms LDA. The precision and recall of this approach are both high across all datasets, approaching 100%. This approach can also extract sentiment variations for every topic.

The performance of this proposed work TTSM using BIE scheme is compared with twoexisting approaches RCB-LDA and LDA. Figure 6.1 shows the performance comparison of the proposed method with other existing approaches based on the four different metrics classification delay, time, processing delay, number of iterations.

Table 6.2 Comparison table

Metrics	RCB-LDA	LDA	Proposed TTSM
Filtering Accuracy	89	90	94
Detection time	73.71	70.68	49.69
Classification Delay	3702.09	2984.06	2108.08
Number of iterations	64.52	57.81	48.21

Here, measure the performance of the existing LDA then measure the results of the TTSMbased classification algorithm. Classification is evaluated by comparing the twitter contents' assigned labels with their true labels provided by the twitter security corpus. Performance comparison of proposed TTSM using BIE with existing approaches based on Result accuracy.

Metrics	RCB-LDA	LDA	Proposed TTSM
Detection Accuracy	89	90	94

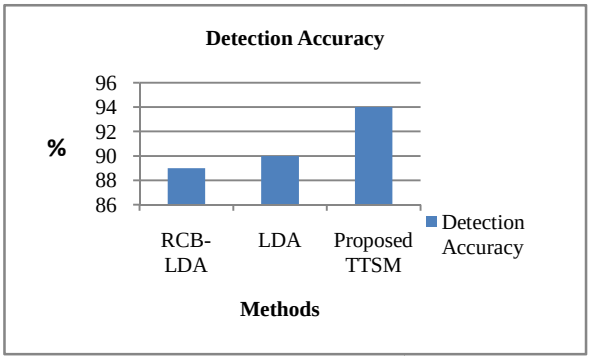

Fig 6.2 Detection Accuracy

From the chart Figure 6.2 shows the performance measure based on the accuracy of detected cluster and the proposed approach TTSM took less time while comparing the other methods and the worst based on the accuracy is RCB-LDA method. Performance comparison of proposed TTSM using BIE with existing approaches based on classification delay.

Metrics	RCB-LDA	LDA	Proposed TTSM
Classification Delay	3702.09	2984.06	2108.08

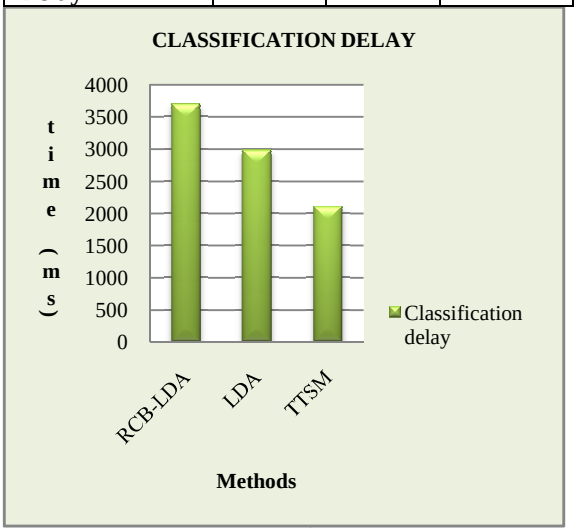

Fig6.3 Classification Delay

From the chart Figure6.3 shows the performance measure based on the classification delay and the proposed approach TTSM took less time while comparing the other methods and the worst time complexity is RCB-LDA method. Performance comparison of proposed TTSM using BIE with existing approaches based on number of iterations.

Metrics	RCB-LDA	LDA	Proposed TTSM
Number of iterations	64.52	57.81	48.21

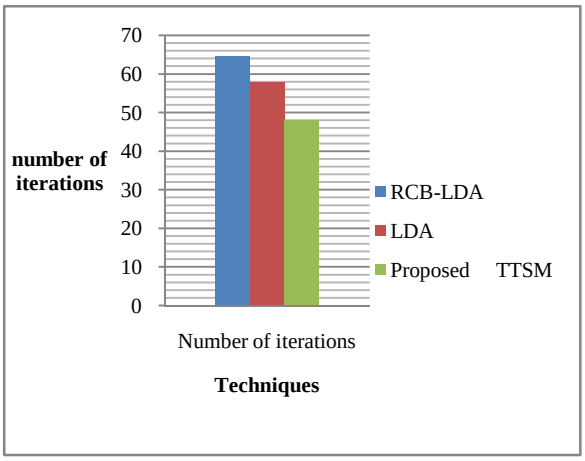

Fig6.4 Comparison of Iterations Execution

From the chart Figure 6.4 shows the performance measure based on the processing of the proposed approach TTSM took less number of iterations for the amount of datasets while comparing the existing method and the worst based on the iterations is RCB-LDA. Performance comparison of proposed TTSM using BIE with existing approaches based on time.

7. CONCLUSION

The proposed work handled the problem of analyzing public sentiment variations and finding the comparison between different topics. To solve the problem, the system proposed two text mining based models, TTSM (Temporal Topic Sentiment Mining) and BIE (Bayesian Information Extractor). The TTSM model can filter unwanted topics and then extract topics and opinions to reveal the variations of sentiment. To give a more intuitive representation, the BIE model can find the best matching keyword in each domain in natural language to provide sentiment variations. The proposed models were evaluated on Twitter dataset Experimental results showed that the proposed models can mine opinions and sentiment variations. Moreover, the proposed models are general. They can be used to discover special topics or aspects in one text collection in comparison with another text collection.In future, TTSM can be enhanced with real-time twitter domain, where every user can view the opinion on the received tweets. This work can also be extended with some other evolutionary algorithm, which can produce better results over dynamic undefined dataset. The main weakness of text mining is in effective review, which means if user uses short messages or local language based tweets then the system needs additional training on it.

REFERENCES

[1] Sil and A. Yates, “Re-ranking for joint named-entity recognition and linking,” in Proc. 22nd ACM Int. Conf. Inf. Knowl. Manage,2013, pp. 2369–2374.

[2] Goldman, Roy, et al. "A standard textual interchange format for the object exchange model (oem)", (1998).

[3] Han, Jiawei, MichelineKamber, and Jian Pei. *Data mining: concepts and techniques: concepts and techniques*. Elsevier, 2011.

[4] J. Gao, M. Li, C. Huang, and A. Wu, “Chinese word segmentation and named entity recognition: A pragmatic approach,” in Comput. Linguist., vol. 31,2005, pp. 531–574.

[5] Jo, Yohan, and Alice H. Oh. "Aspect and sentiment unification model for online review analysis." *Proceedings of the fourth ACM international conference on Web search and data mining*. ACM, 2011.

[6] Kennedy, Alistair, and Diana Inkpen. "Sentiment classification of movie reviews using contextual valence shifters". *Computational intelligence* 22.2 ,2006, pp.110-125.

[7] Kulkarni, Anagha, and Ted Pedersen. "Name Discrimination and Email Clustering using Unsupervised Clustering and Labeling of Similar Contexts."*IICAI*. 2005

[8] Lin, Chenghua, et al. "Weakly supervised joint sentiment-topic detection from text." *Knowledge and Data Engineering, IEEE Transactions on* 24.6, 2012,pp:1134-1145.

[9] Pang, Bo, and Lillian Lee. "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts." *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2004.

[10] W. Jiang, L. Huang, and Q. Liu, “Automatic adaption of annotation standards: Chinese word segmentation and pos tagging- A case study,” in Proc. Joint Conf. 47th Annu. Meeting ACL 4th Int. Joint Conf. Natural Language Process. AFNLP, 2009, pp. 522–530.

[11] X. Wang, A. McCallum, and X. Wei, “Topical n-grams: Phrase and topic discovery, with an application to information retrieval,” in Proc. IEEE 7th Int. Conf. Data Mining, 2007, pp. 697–702.

[12] Yang, Ming, et al. "Social media analytics for radical opinion mining in hate group web forums." *Journal of homeland security and emergency management* 8.1,2011.

[13] Y. Zhang and S. Clark, “A fast decoder for joint word segmentation and pos-tagging using a single discriminative model,” in Proc. Conf. Empirical Methods Natural Language Process., 2010, pp.843–852.